Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI

Year: 2020 | Citations: 690 | Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Ádám Nagy, M. Srikumar

Abstract

The rapid spread of artificial intelligence (AI) systems has precipitated a rise in ethical and human rights-based frameworks intended to guide the development and use of these technologies. Despite the proliferation of these "AI principles," there has been little scholarly focus on understanding these efforts either individually or as contextualized within an expanding universe of principles with discernible trends.

To that end, this white paper and its associated data visualization compare the contents of thirty-six prominent AI principles documents side-by-side. This effort uncovered a growing consensus around eight key thematic trends: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. Underlying this “normative core,” our analysis examined the forty-seven individual principles that make up the themes, detailing notable similarities and differences in interpretation found across the documents. In sharing these observations, it is our hope that policymakers, advocates, scholars, and others working to maximize the benefits and minimize the harms of AI will be better positioned to build on existing efforts and to push the fractured, global conversation on the future of AI toward consensus.